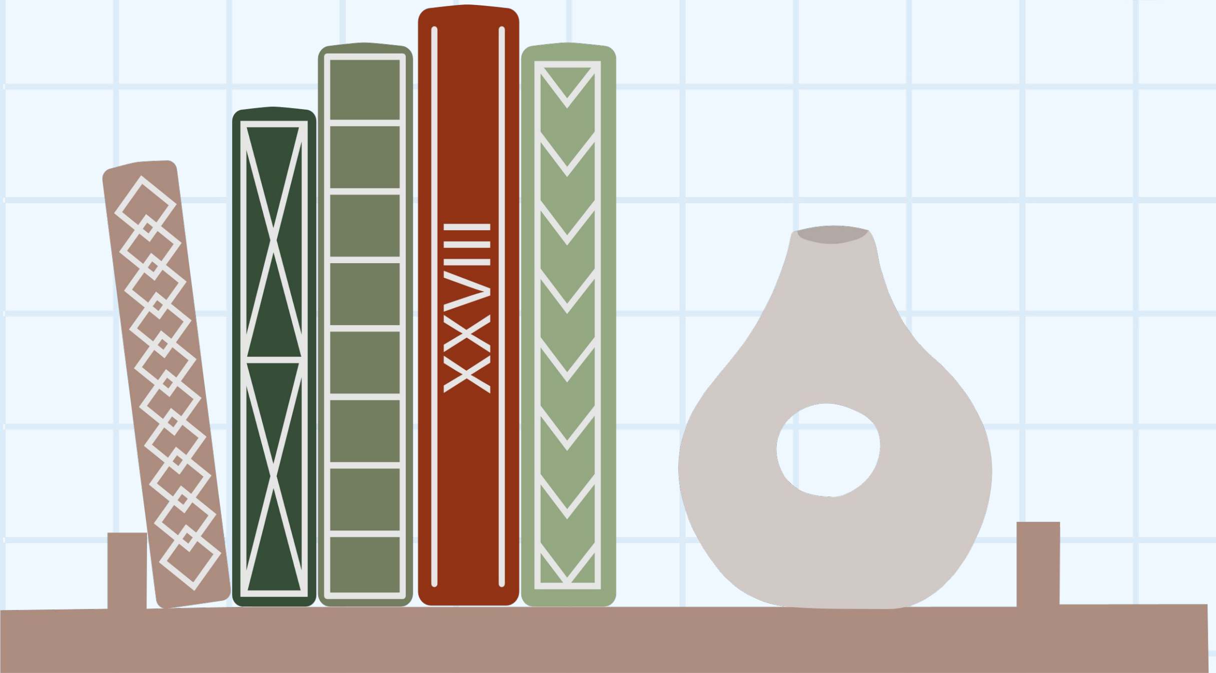
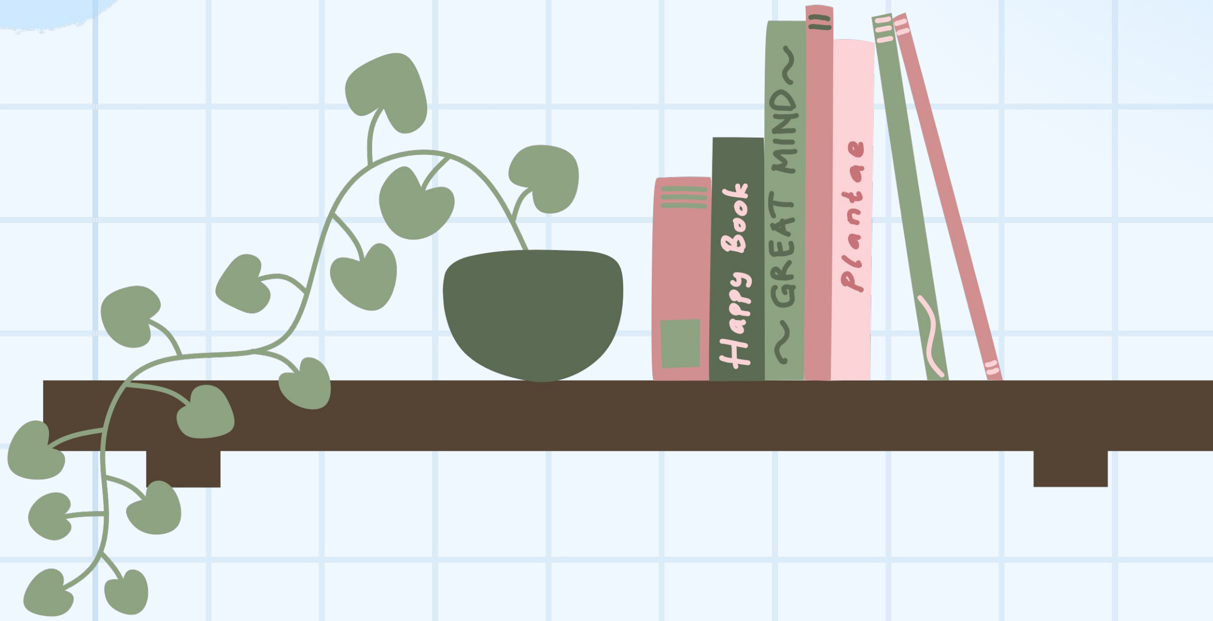
Powered by
Futurense

BS./BSC.IN

Applied AI and Data Science

Basics of Data Analytics





Let's dive into and learn:

1 Cleaning Data

Cleaning Data

- Recall the different steps of Data Analytics
- After the data gathering, comes the step of cleaning the data
- Before cleaning, data may have issues (errors etc.)
- Without cleaning, the errors in your data might affect your exploration and analysis
- Can lead you to drawing the wrong conclusions

Cleaning Data

- Data Cleaning workflow includes the steps of
  - Inspecting Data
  - Cleaning Data
  - Verification of Data

Inspecting Data

- Data Inspection includes
  - Detecting issues and errors
  - Comparing against rules and constraints
  - Data Profiling
  - Visualizing data

Inspecting Data

- Data Inspection includes
  - Looking at the source data to understand the structure
  - Contents of the data
  - Relations between data to uncover any anomalies or data quality issues

Cleaning Data

- Data cleaning usually involves the following steps
    - Removing duplicates
    - Formatting data
    - Missing values
    - Errors or discrepancies

Tools for Data Cleaning

- Excel / Spreadsheets
- Google DataPrep
- Python
- R

Excel / Spreadsheets

- Many built-in formulae and features that help identify issues, clean, format and transform data
- Additional add-ins available for advanced features
- Add-ins to import data from several different types of sources —such as Microsoft Power Query for Excel and Google Sheets Query function for Google Sheets.

Google Data Prep

- An intelligent cloud data service
- Allows us to explore, clean, and prepare both structured and unstructured data for analysis
- Easy to use
- Can automatically detect schemas, data types, and anomalies

Python

- Python is a programming language useful for data cleaning
- Has a large library of packages that allow data cleaning
- Jupiter Notebook - open-source web application used for data cleaning and modelling
- Numpy - fast, versatile, interoperable, and easy to use. Can work with large data
- Pandas - fast and easy data analysis operations, complex operations such as merging, joining, and transforming huge chunks of data, prevent common errors that result from misaligned data coming in from different sources.

R

- R is a programming language useful for statistical computing
- Has a large library of packages that allow data cleaning
- Dplyr - library for data wrangling with straightforward syntax
- Data.table - helps to aggregate large data sets quickly.
- Jsonlite - is a robust JSON parsing tool, great for interacting with web APIs.

Verification

- Inspect the results to establish effectiveness of the data cleaning operation.
- Document all changes to the data through the data cleaning process
- Reasons for the changes
- Quality of the currently stored data

Recap

Powered by

Futurense

Steps of Data cleaning

Powered by

Futurense

Thank you